Consensus for partial trees

Alain Guénoche

CNRS, Institut de Mathématiques de Luminy
guenoche@iml.univ-mrs.fr

Jobim 2013 ?

Pre-requisite

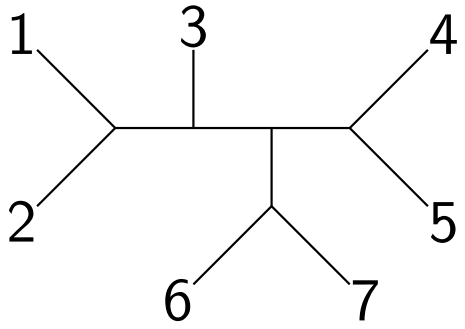
A X-tree is :

- an unrooted tree,
- X is the set of n leaves,
- nodes have degree at least 3,
- edges have a positive length.

X-tree $\Longleftrightarrow$ { bipartitions } $\Longleftrightarrow$ Tree distance

- n external edges (to leaves) common to every X-tree
- at most $n-3$ internal edges making splits
- $D_U(x,y) =$ Nb. of edges along the path between x and y

An X-tree and two equivalent representations



D_U	1	2	3	4	5	6	
2	2						
3	3	3					1 2 | 3 4 5 6 7
4	5	5	4				1 2 3 | 4 5 6 7
5	5	5	4	2			1 2 3 4 5 | 6 7
6	5	5	4	4	5		1 2 3 6 7 | 4 5
7	5	5	4	4	4	2	

Unitary distance

Bipartitions

Two Consensus Trees

$\Pi = \{T_1, \ldots, T_m\}$ a *profile* of m X-trees

A consensus tree C is a X-tree *summarizing* Π

- Majority consensus tree (MCT) (Mc Morris, 1983)
  - MCT C is median for the *Robinson-Foulds* distance

$$\sum_{i=1}^{m} D_{R-F}(C, T_i) \text{ minimum}$$

  - Computable in $O(nm)$ (from the split table)
  - Minority edges are not significant in evolution
- Average Consensus Tree (Lapointe & Cucumel, 1997)
  - Adding the m unitary distances (which does not make a tree distance),
  - Apply a distance method (NJ)

The consensus tree quality criteria

$\{P_1, \ldots P_q\}$ majority bipartitions in the m trees

- Rate of majority bipartitions

$$\tau_{maj} = \frac{|Majority\ \ bipartitions|}{|bipartitions|}$$

- Consensus tree weight

$$W_{\Pi}(C) = \sum_{P_k \in C} |\{T_i \in \Pi\ \ containing\ \ P_k\}|$$

- Normalized weight

$$\mathcal{W}(C) = \frac{W_{\Pi}(C)}{m \times (n-3)}$$

First comparisons

- A model tree with 16 taxa
- 30 swapped trees (exchanging 2 leaves in τ_a % of the trees)
- Computing the two consensus trees, comparing them to the model tree
  - *Id* the percentage of identity between the model tree and the consensus tree
  - $\overline{RF}$ the average value of the Robinson-Foulds distance

	Median tree		Average tree	
τ_a	*Id*	$\overline{RF}$	*Id*	$\overline{RF}$
50%	.97	.06	1.0	0.0
75%	.76	.84	.99	.02
100%	.14	4.24	.99	.02

When the X sets are not identical ?

Because

- genomes are not completely sequenced
- some genes have disappeared
- orthologous gene cannot be identified

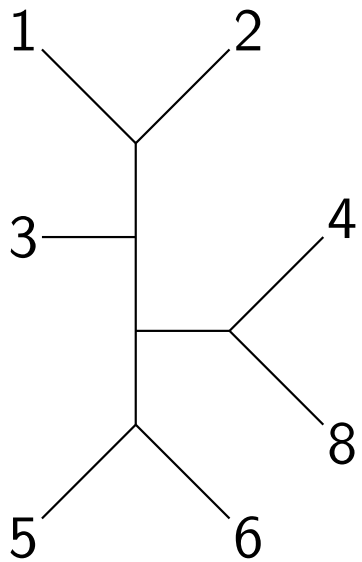
One solution : to extend the partial trees

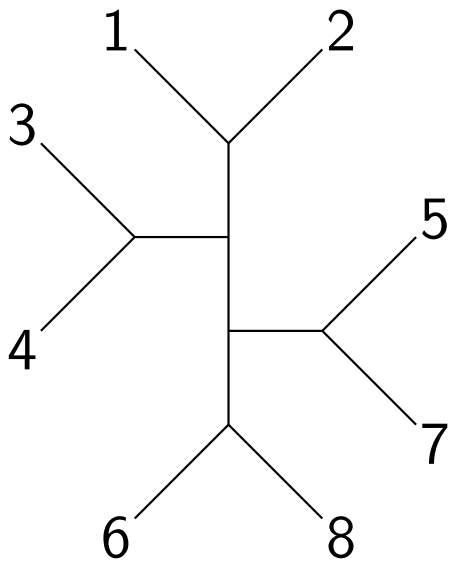
- inferring the missing class numbers
- according to the closest bipartitions

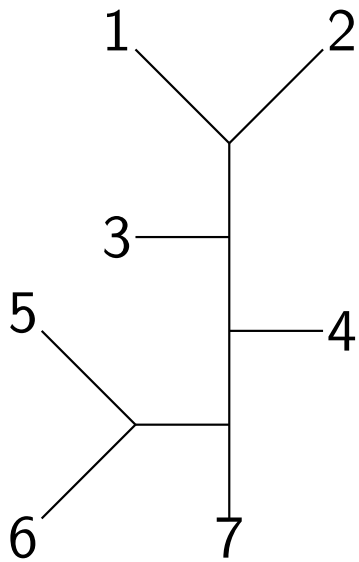
Another solution

- computing the average unitary distance between taxa (if all the pairs are present at least one time)
- building the NJ tree

A small example







Partitions to be completed

$$P(x) =?, Q(x) \neq? \Rightarrow D(P,Q) = \frac{|y \in Y \text{ such that } P(y) \neq Q(y)|}{|y \text{ tel que } P(y) \neq? \text{ and } Q(y) \neq?|}$$

	1	2	3	4	5	6	7	8	C_{bip}	Dmin	IC
1	1	1	0	0	0	0		0	5,9	0	0
2	1	1	1	0	0	0		0	10	0	0
3	1	1	1	0	1	1		0	8	2/7	1
4	1	1	1	1	0	0		1	11	0	0
5	1	1	0	0	0	0	0	0			
6	1	1	1	1	0	0	0	0			
7	1	1	1	1	0	1	0	1			
8	1	1	1	1	1	0	1	0			
9	1	1	0	0	0	0	0		1,5	0	0
10	1	1	1	0	0	0	0		2	0	0
11	1	1	1	1	0	0	0		6	0	0
12	1	1	1	1	0	0	1		4	0	1

- C_{bip} : closest bipartitions,
- IC : Inferred Cluster = majority class number in C_{bip}

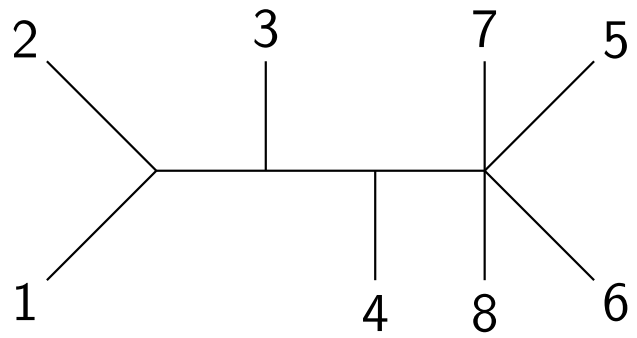
Averaging the unitary distances

	1	2	3	4	5	6	7	8
1	0	2	3.33	4.33	5.33	5.33	5	5
2	2	0	3.33	4.33	5.33	5.33	5	5
3	10/3	10/3	0	3	4.66	4.66	4.5	4.5
4	13/3	13/3	3	0	4.33	4.33	4	3.5
5	16/3	16/3	14/3	13/3	0	2.66	2.5	4
6	16/3	16/3	14/3	13/3	8/3	0	3.5	3
7	5	5	9/2	4	5/2	7/2	0	4
8	5	5	9/2	7/2	4	3	4	0

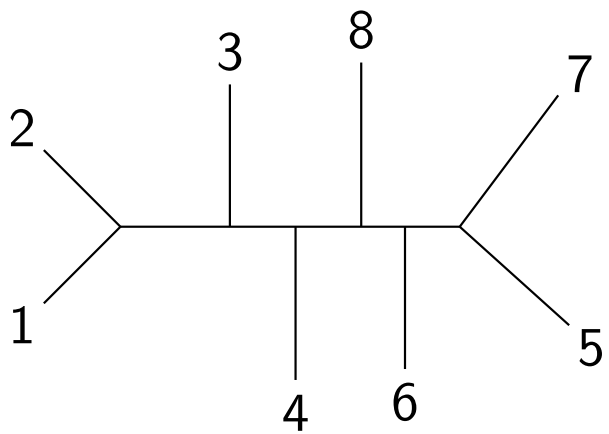
Average values of the unitary distances
in ratio form (lower-left) and decimal form (upper right-right)

Two consensus trees

Median Consensus Tree



Average Consensus Tree



Majority/Average tree

Distance methods $\Rightarrow$ fully resolved trees $\Rightarrow$ minority edges

- Comparing Average consensus tree to each tree in the profile only considering common taxa
- Keeping majority edges

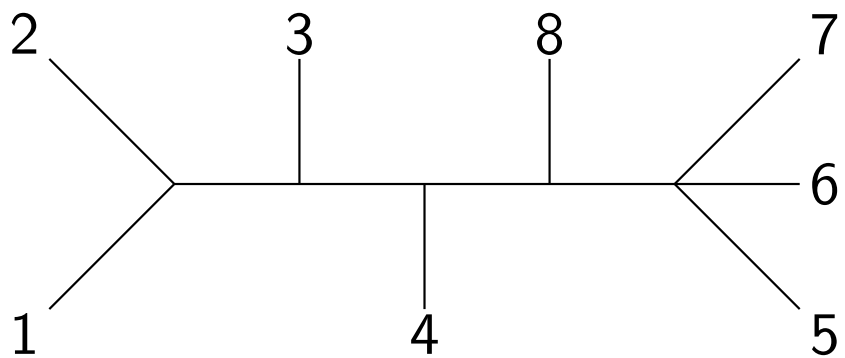


Figure: Majority/Average tree

First simulations : swapped trees

30 trees on $|X| = 16$ taxa

- Random rooted topology on X (Yule-Harding)
- swapped trees, exchanging 2 taxa in τ_a % of the trees
- C_{med} = Median consensus of the complete trees

30 partial trees (on $X' \subset X$)

1. erasing 1 random taxon in each tree
2. computing Majority, Average and Average/Majority consensus of partial trees

Swapped trees

Average values over 100 trials

	C_{med}		Majority			Average		Ave/Maj	
τ_a	τ_{maj}	$\mathcal{W}$	τ_{err}	Id	RF	Id	RF	Id	RF
50%	1.0	.83	.059	.94	.06	.99	0.01	1.0	0.0
75%	.97	.74	.080	.71	.36	.74	0.37	1.0	0.0
100%	.82	.58	.096	.42	.77	.10	2.36	.97	0.3

- τ_{err} = Rate of errors predicting the class number
- Id = Identity rate between C_{med} and consensus trees
- RF = Robinson-Foulds average distance between C_{med} consensus trees

Second simulations : Pruned trees

30 trees on $|X| = 16$ taxa

- One model tree : random rooted topology on 16 taxa
- pruned trees, erasing NbS random leaves in each tree

	Majority		Average		Ave/Maj	
NbS	*Id*	*RF*	*Id*	*RF*	*Id*	*RF*
1	1.0	0.0	1.0	0.0	1.0	0.0
2	.73	.27	1.0	0.0	.97	0.03
3	.07	1.69	1.0	0.0	.58	0.48
4	0.0	3.68	1.0	0.0	.07	1.85

Identity rate and RF distance between model and consensus trees

Real data (C. Brochier)

Complete trees

- 47 *Archeae* (Methanogens)
- 185 gene trees aligned by MAFFT 7 + computed by PhyML 3
  - PhyML 3 of concatenated alignments $\Rightarrow$ 43 majority edges
  - Majority Consensus $\Rightarrow$ 25 majority edges

Partial trees

- for each tree, erase 10% of the sequences in the average (from 3 to 15)
- 185 partial trees computed by MAFFT 7 + PhyML on the remaining genes

ML & Median consensus of complete trees

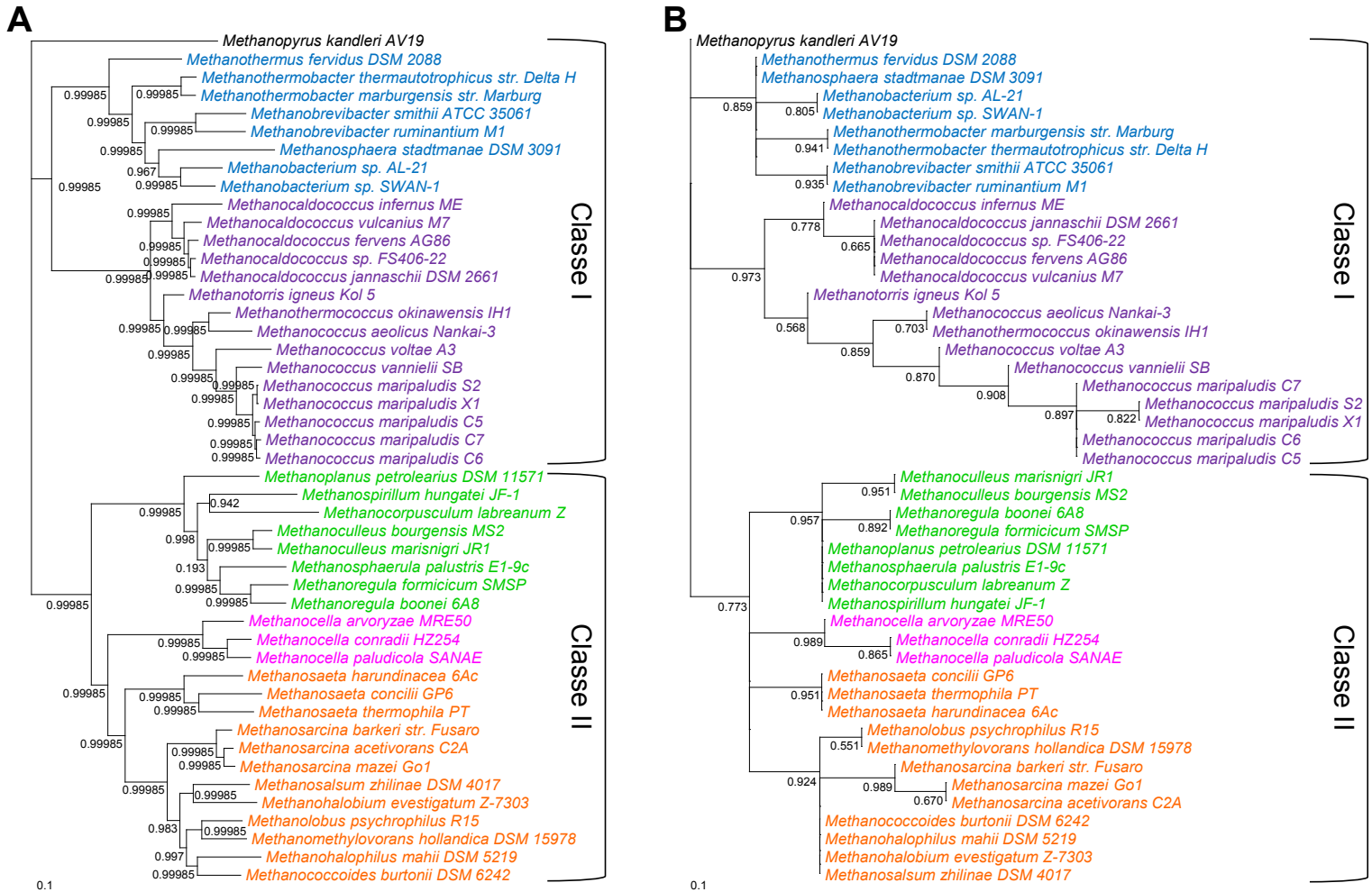


Figure 5. (A) Arbre de maximum de vraisemblance inféré à partir des 185 gènes présents en une seule copie dans les génomes des 47 méthanogènes analysés (46657 positions). **(B)** Arbre consensus majoritaire des 185 arbres complets. Methanopyrales (noir), Methanobacteriales (bleu), Methanococcales (violet), Methanomicrobiales (vert), Methanocellales (rose), Methanosarcinales (orange).

Consensus of 185 partial trees

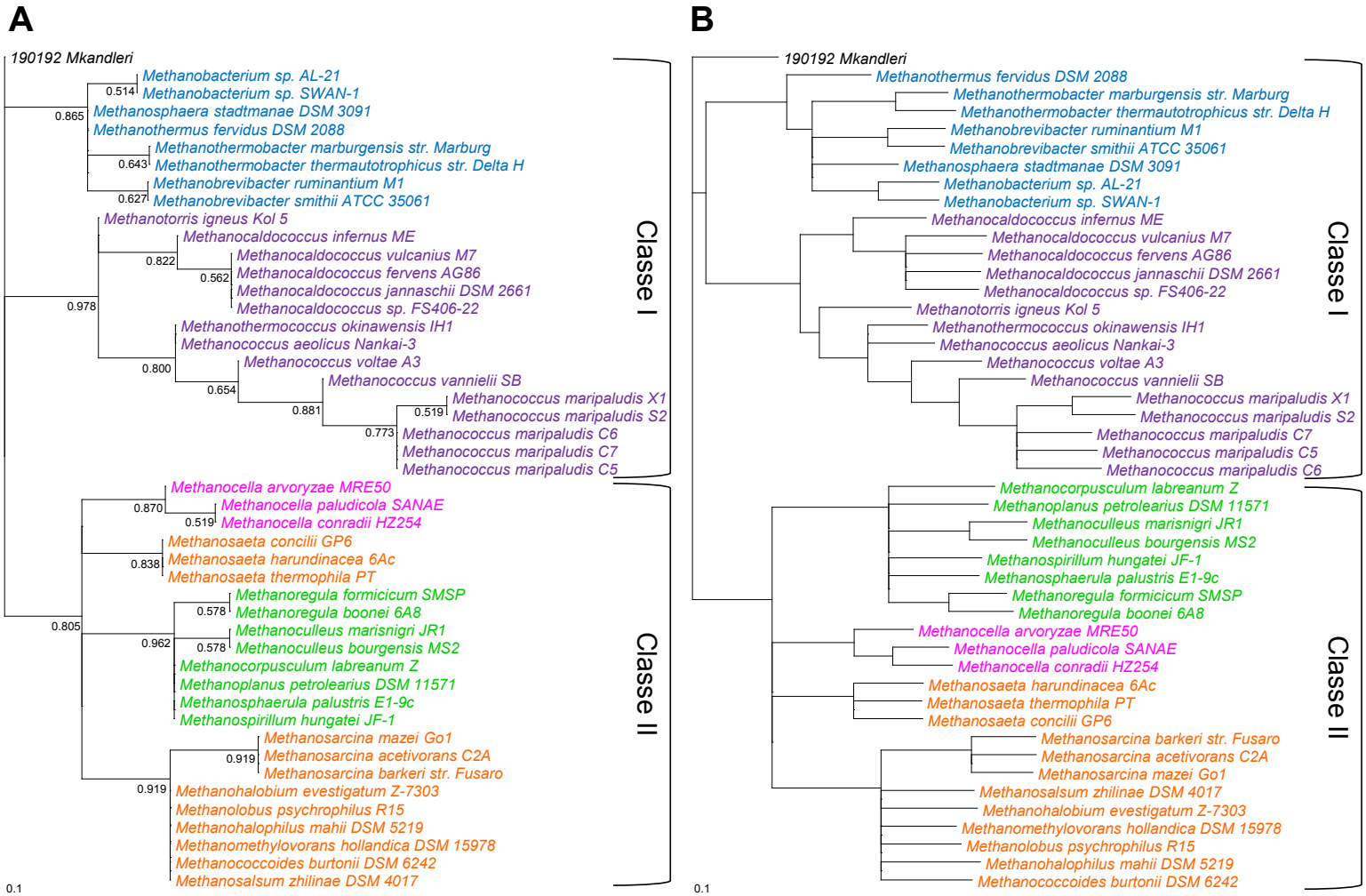


Figure 6. (A) Arbre médian des 185 arbres incomplets. **(B)** Arbre consensus moyen des 185 arbres incomplets. Methanopyrales (noir), Methanobacteriales (bleu), Methanococcales (violet), Methanomicrobiales (vert), Methanocellales (rose), Methanosarcinales (orange).

Conclusions

On swapped trees

- the rate of errors, when predicting classes of missing taxa, is small ($\leq 10\%$)
- the RF distance is smaller for Average/Majority tree which is identical to the Median consensus from complete trees

On pruned trees

- the Median tree becomes poor when the number of suppressed taxa increases
- the Average/Majority tree is much better than the Majority tree

What to do now ?

- to compare partial consensus tree to "super-tree" methods
- to define the Average Consensus tree when distance values are missing
  - Estimating distance values according to a tree model A. Guénoche, S. Grandcolas, Estimating Missing Values in Tree Distances. In *Data Analysis, Classification and Related Methods*, Springer, 143–148, 2000.
    V. Makarenkov, J.-F. Lapointe, A Weighted Least-Squares Approach for Inferring Phylogenies from Incomplete Distance Matrices. *Bioinformatics*, 20, 2113–2121, 2004.
  - Recovering a tree from a partial distance
    A. Guénoche, B. Leclerc, V. Makarenkov, On the extension of a partial metric to a tree metric, *Discrete Maths*, 276/1-3, 229-248, 2004.

Many thanks to ..

- Céline Brochier - Armanet (LBBE, Lyon)
- Vladimir Makarenkov (UQAM, Montréal)
- Laurent Tichit (IML, Marseille)
-
- LIRCO (Laboratoire International Franco-Québécois de Recherche en Combinatoire)